

Introduction to Bioinformatics

2. DNA Sequence Retrieval and comparison

Benjamin F. Matthews
United States Department of Agriculture
Soybean Genomics and Improvement
Laboratory
Beltsville, MD 20708
matthewb@ba.ars.usda.gov

What we will cover today

- ✠ Retrieving a known DNA sequence
- ✠ Similarity searching with a DNA sequence
- ✠ BLAST

—•—•—•—•—•—•—

✪ Is at least some of the DNA sequence available? (EST sequence)?

Finding Sequences in Databases

- ✠ The public DNA and protein sequence databases are huge.
- ✠ In order for these databases to be useful, the data must be readily accessible to researchers.

What Are You Looking For?

✠ A gene?

- ◆ DNA or protein sequence?
- ✠ DNA sequences are essentially all in **GenBank**
 - ◆ Genomic, mRNA, cDNA, EST?
- ✠ Proteins are harder to pin down
 - ◆ **GenPept** (GenBank Peptides) is huge and poorly annotated - lots of junk
 - ◆ **SwissProt** is carefully annotated, but not fully comprehensive
 - ◆ **PIR** is somewhere in between

Large Databases

- ✧ Once upon a time, **GenBank** sent out sequence updates on CD-ROM disks a few times per year.
- ✧ Now **GenBank** is over 95 Gigabytes (28 **billion** bases)
- ✧ Most biocomputing sites update their copy of **GenBank** every day over the internet.
- ✧ Scientists access **GenBank** directly over the Web

You can search DNA sequence database

- ✧ Retrieve known sequences by
 - ◆ Keyword search
 - ◆ Accession numbers
- ✧ If you know some DNA sequence
 - ◆ Compare your DNA sequence with those in database
 - ◆ Basic Local Alignment Search Tool (BLAST) searches

Retrieve a DNA sequence

✠ ENTREZ

◆ <http://www.ncbi.nlm.nih.gov/Entrez/>

✠ Click – Nucleotide

◆ GenBank

◆ OR

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide>

✠ Accession number

✠ Keyword search

Entrez is a Tool for Finding Sequences

✠ **GenBank** is managed by the **NCBI** (National Center for Biotechnology Information) which is a part of the US National Library of Medicine.

✠ NCBI has created a Web-based tool called **Entrez** for finding sequences in **GenBank**.

<http://www.ncbi.nlm.nih.gov>

✠ Each sequence in **GenBank** has a unique “**accession number**”.

✠ **Entrez** can also search for keywords such as gene names, protein names, and the names of organisms or biological functions

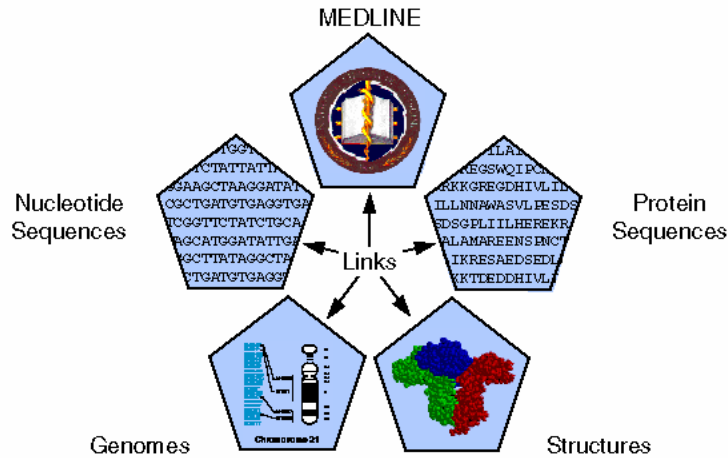
Entrez is a Database

- ✠ The **Entrez** database contains all of the nucleotide and protein sequences in **GenBank** (updated daily) along with all of the literature in **MEDLINE** and the 3-D protein structures in **PDB** (**Protein Data Base**).
- ✠ **Entrez** is much more than a database, it is both a powerful search engine and a pre-computed list of relationships among all of its data elements

Entrez is Internally Cross-linked

- ✠ DNA and protein sequences are linked to other similar sequences
- ✠ **Medline** citations are linked to other citations that contain similar keywords
- ✠ 3-D structures are linked to similar structures

Databases contain more than just DNA & protein sequences



The screenshot shows the NCBI Entrez search engine homepage. The header includes the NCBI logo and the text 'Entrez, The Life Sciences Search Engine'. Below the header is a navigation bar with links to 'HOME', 'SEARCH', 'SITE MAP', 'PubMed', 'Entrez', 'Human Genome', 'GenBank', 'Map Viewer', and 'BLAST'. A search bar is located below the navigation bar, with the text 'Search across databases' and a 'GO' button. The main content area is titled 'Welcome to the new Entrez cross-database search page' and contains a grid of database links. A red arrow points to the 'Nucleotide' database link in the first column of the grid.

Welcome to the new Entrez cross-database search page	
<ul style="list-style-type: none"> PubMed: biomedical literature citations and abstracts PubMed Central: free, full text journal articles 	<ul style="list-style-type: none"> Books: online books OMIM: Online Mendelian Inheritance in Man Site Search: NCBI web and FTP sites
<ul style="list-style-type: none"> Nucleotide: sequence database (GenBank) Protein: sequence database Genome: whole genome sequences Structure: three-dimensional macromolecular structures Taxonomy: organisms in GenBank SNP: single nucleotide polymorphism Gene: gene-centered information 	<ul style="list-style-type: none"> UniGene: gene-oriented clusters of transcript sequences CDD: conserved protein domain database 3D Domains: domains from Entrez Structure UniSTS: markers and mapping data PopSet: population study data sets GEO: expression and molecular abundance profiles GEO Datasets: experimental sets of GEO data
<ul style="list-style-type: none"> Journals: detailed information about the journals indexed in PubMed and other Entrez databases 	<ul style="list-style-type: none"> MeSH: detailed information about NLM's controlled vocabulary

Enter terms and click 'GO' to run the search against ALL the databases, OR
Click Database Name or icon to go directly to the Search Page for that database, OR
Click Question Mark for a short explanation of that database.

GenBank

✧ National Institute of Health, National Library of Medicine, National Center for Biotechnology Information

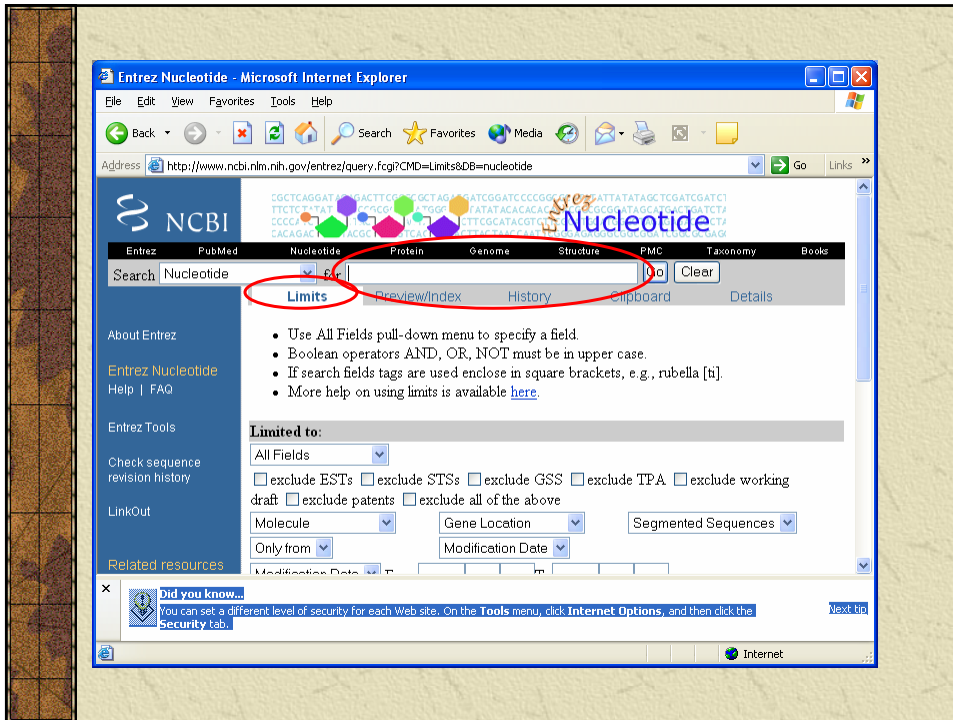
✧ <http://www.ncbi.nlm.nih.gov/>

✧ <http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html>

✧ Retrieve a sequence from GenBank

✧ Analyze raw sequence data

- ◆ Base calling
- ◆ Editing
- ◆ Obtaining a consensus sequence
- ◆ Translating
- ◆ Restriction mapping
- ◆ Similarity comparisons
- ◆ Motif searches

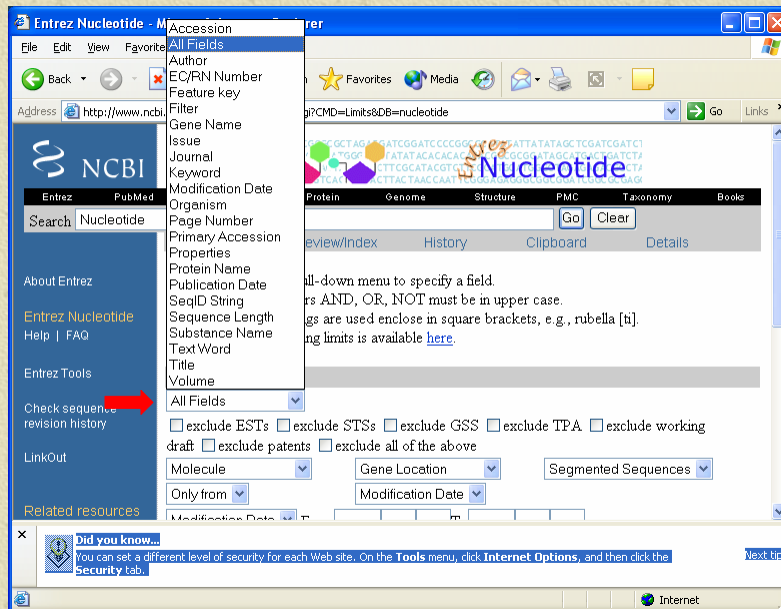
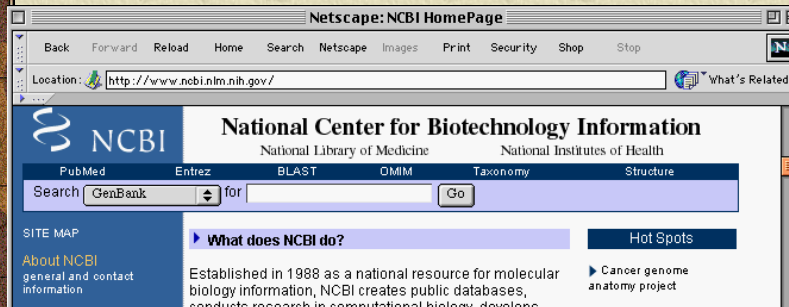


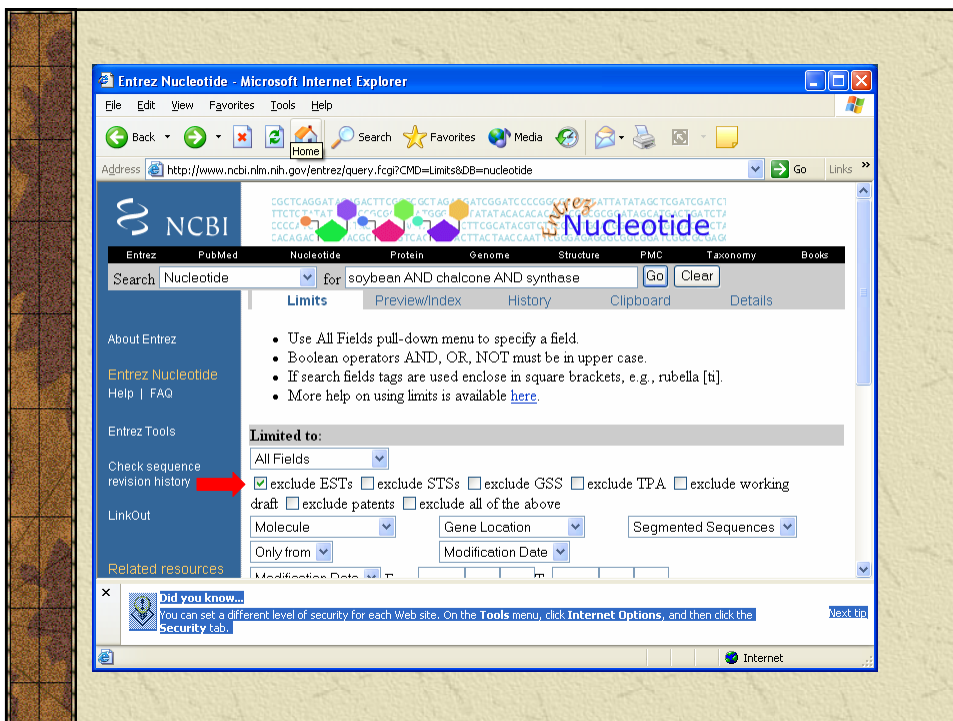
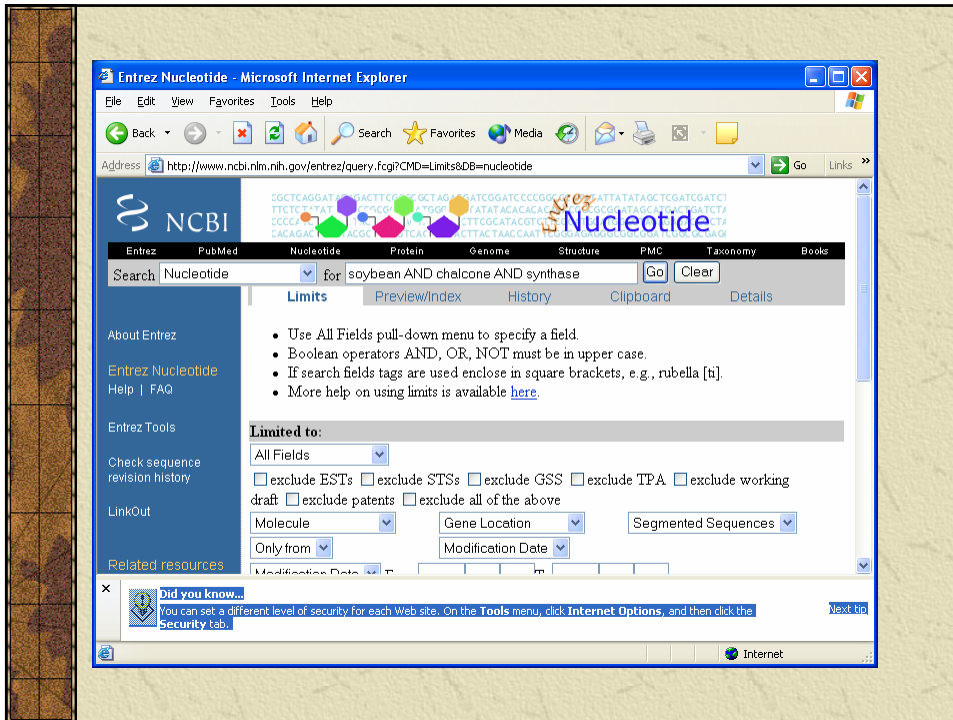
Accession Numbers!!

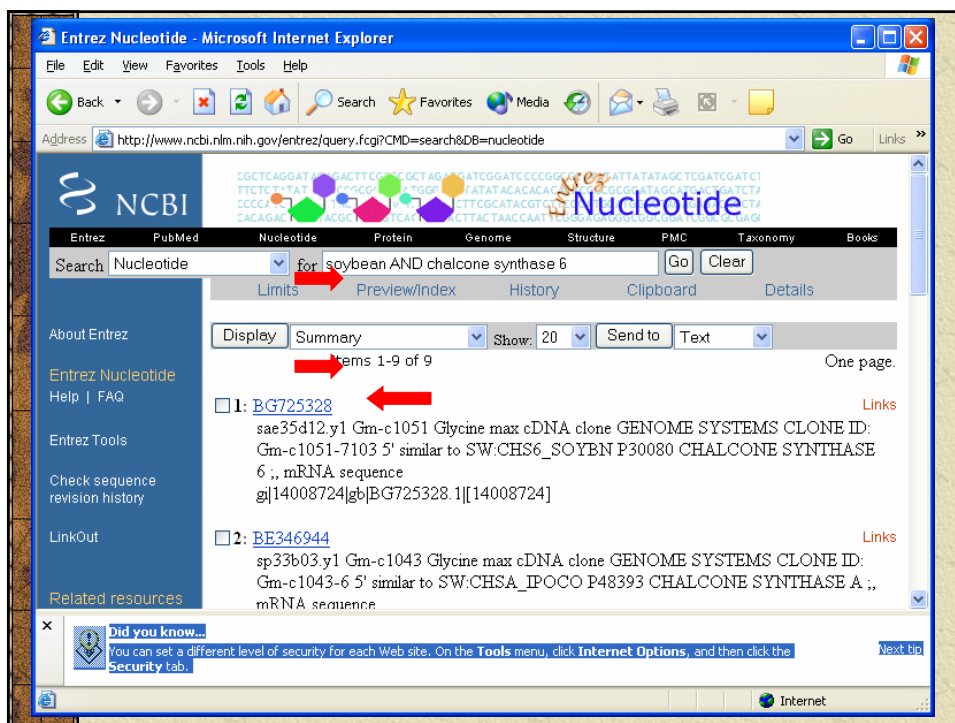
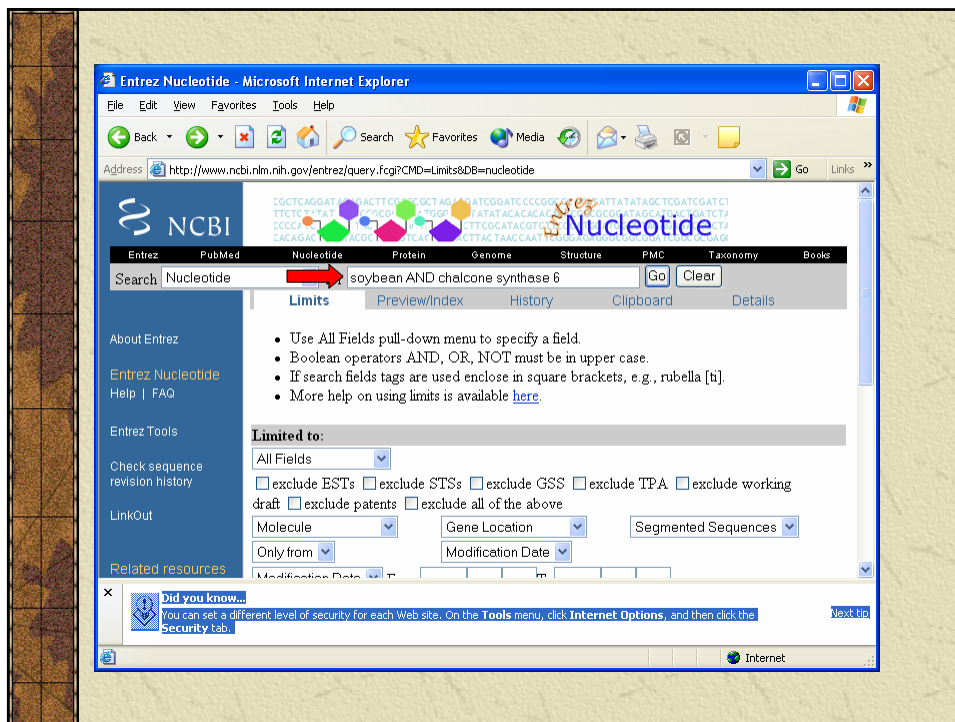
- ✧ Databases are designed to be searched by accession numbers (and locus IDs)
- ✧ These are guaranteed to be non-redundant, accurate, and not to change.
- ✧ Searching by gene names and keywords is inexact and retrieves more than one record usually

Type in a Query term

✳ Enter your search words in the query box and hit the “Go” button







NCBI Sequence Viewer - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=14008724> Go Links

NCBI Nucleotide

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search Nucleotide for [] Go Clear

Limits Preview/Index History Clipboard Details

Display default Show: 20 Send to File Links

1: [BG725328](#) sae35d12.y1 Gm-c1...[gi:14008724]

IDENTIFIERS

dbEST Id: 8490125
EST name: sae35d12.y1
GenBank Acc: BG725328
GenBank gi: 14008724

CLONE INFO
Clone Id: GENOME SYSTEMS CLONE ID: Gm-c1051-7103 (5')
DNA type: cDNA

PRIMERS
PolyA Tail: Unknown

SEQUENCE

```

GGTTGTGCCAAGTCCATTTTCTATTGACTTCTTCTCAATTTGATCCAAAGATGAACAGCAC
ACATGCACTTGACATGTTACCATACTCGCTAAGCACGTGTCTAGTAGCTTCCATTTTTC
ATGCTTCAATCCTAACTTAGCCTCAACTTGGTCCAAAATTTGCTGGTCCACAGGGTGTGC
AATCCAAAAGATAGAGTTGTAATCATCAATTTCTAAGGGTTTGAAGGCTTCAACCAAGGC
CTTTTCGATGTTCTTGGAGATGAGTCCAGGAACATCCTTGAAGGATGGAAGTGAAGTCC

```

Did you know...
You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.

NCBI Sequence Viewer - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=14008724> Go Links

SEQUENCE

```

GGTTGTGCCAAGTCCATTTTCTATTGACTTCTTCTCAATTTGATCCAAAGATGAACAGCAC
ACATGCACTTGACATGTTACCATACTCGCTAAGCACGTGTCTAGTAGCTTCCATTTTTC
ATGCTTCAATCCTAACTTAGCCTCAACTTGGTCCAAAATTTGCTGGTCCACAGGGTGTGC
AATCCAAAAGATAGAGTTGTAATCATCAATTTCTAAGGGTTTGAAGGCTTCAACCAAGGC
CTTTTCGATGTTCTTGGAGATGAGTCCAGGAACATCCTTGAAGGATGGAAGTGAAGTCC
TACTTGGCGAAGGTGGCCATCAATAGCGCCTTTCGCTGTCTGGAAGGATTTGTTGTCAGT
CCACACAAGCTCAAAACAAAGGCTTTTCAGCTGGCAGAGGATCTGATCCAAACATGACAGC
AGCTGCACCATCTCCAAACAAAGGCTTGCCCCACAGGCTGTCAAGATGTGTGCTACTCGG
GCCACGAAATGTGACTGCTGTGATCTCCGA

```

Quality: High quality sequence stops at base: 400

Entry Created: May 8 2001
Last Updated: Jul 22 2004

COMMENTS

When it has been determined, an EST from the other end of this clone is listed in the 'Other ESTs on clone' field. Possible reversed clone: similarity on wrong strand This clone is available through: Biogenetic Services, 801 32nd Ave. Brookings, SD 57006 USA (phone: 800 423 4163; email: info@biogeneticservices.com)

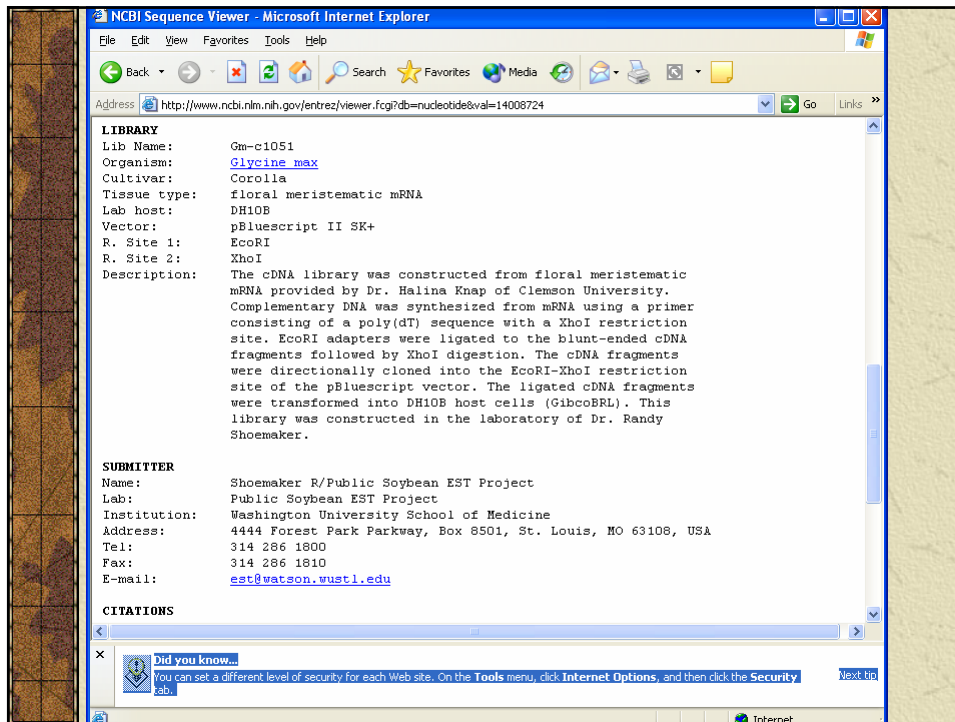
PUTATIVE ID

Assigned by submitter
SW:CHS6_SOYBN P30080 CHALCONE SYNTHASE 6 ;

LIBRARY

Lib Name: Gm-c1051
Organism: [Glycine max](#)
Cultivar: Corolla
Tissue type: floral meristematic mRNA
Isb host: pUC19

Did you know...
You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab.



GenBank Records

- ✠ Databases are composed of records
- ✠ Flat File Format
- ✠ Provides information
- ✠ Standard, consistent organization of data

Flat file format

- ✧ Organized in a structured manner
- ✧ One big file
- ✧ Large body of information assembled and distributed in consistent format
- ✧ Lack support for procession transactions (inserts and updates)

The screenshot shows a web browser window displaying the NCBI Sample GenBank Record page. The browser's address bar shows the URL: <http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html#GenBankDivisionB>. The page features the NCBI logo and navigation tabs for PubMed, Entrez, BLAST, OMIM, Taxonomy, and Structure. The main heading is "GenBank Flat File Format", followed by a paragraph explaining that clicking on any link in the sample record provides a detailed description of that data element or field, and that all descriptions are included on this page, which can be printed as a single document. Below this, the flat file format data is displayed in a structured manner, including fields for Locus, Definition, Accession, Version, Keywords, Source, Organism, Reference, Authors, Title, Journal, Medline, and Reference. The data is organized into sections for each field, with sub-sections for each entry.

NCBI Sample GenBank Record

PubMed Entrez BLAST OMIM Taxonomy Structure

GenBank Flat File Format

Click on any link in this sample record to see a detailed description of that data element or field. All of the descriptions are included on this page, so it can be printed as a single document. You can also return to the [Alphabetical Quicklinks Table](#) or [Resource Guide](#)

LOCUS SCU49845 5028 bp DNA PLN 21-JUN-1999
DEFINITION Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION U49845
VERSION U49845.1 GI:1293613
KEYWORDS
SOURCE baker's yeast.
ORGANISM Saccharomyces cerevisiae
Eukaryota; Fungi; Ascomycota; Hemiascomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE
1 (bases 1 to 5028)
AUTHORS Torpey, L.E., Gibbs, P.E., Nelson, J. and Lawrence, C.W.
TITLE Cloning and sequence of REV7, a gene whose function is required for DNA damage-induced mutagenesis in Saccharomyces cerevisiae
JOURNAL Yeast 10 (11), 1503-1509 (1994)
MEDLINE 95176709
REFERENCE
2 (bases 1 to 5028)
AUTHORS Roemer, T., Madden, K., Chang, J. and Snyder, M.
TITLE Selection of axial growth sites in yeast requires Axl2p, a novel plasma membrane glycoprotein
JOURNAL Genes Dev. 10 (7), 777-793 (1996)
MEDLINE 96194260
REFERENCE
3 (bases 1 to 5028)

Some Fields of GenBank Record

- ✧ Locus Name
- ✧ Sequence length
- ✧ Molecule type
- ✧ Definition
- ✧ GenBank accession number
- ✧ Version
- ✧ Keywords
- ✧ Source
- ✧ Organism
- ✧ Reference
- ✧ Reference
- ✧ Authors
- ✧ Title
- ✧ Journal
- ✧ Medline
- ✧ Other references
- ✧ features
- ✧ Amino acid translation
- ✧ Nucleotide sequence

The screenshot shows a web browser window displaying a GenBank record. The browser's address bar shows the URL: <http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html#GenBankDivisionB>. The page title is "NCBI Sample GenBank Record". Below the title, there are navigation links: PubMed, Entrez, BLAST, OMIM, Taxonomy, and Structure. The main content area is titled "GenBank Flat File Format" and includes a paragraph explaining that clicking on any link in the sample record will show a detailed description of that data element or field. Below this, the record data is displayed in a flat file format. Several fields are circled in red: Locus, Definition, Accession, Version, Keywords, Source, Organism, Reference, Authors, Title, Journal, and Medline. The record data is as follows:

```

LOCUS       SCU49845               5028 bp    DNA             PLN             21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1               GI:1293613
KEYWORDS
SOURCE      baker's yeast.
ORGANISM    Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Hemiascomycetes; Saccharomycetales;
            Saccharomycetaceae; Saccharomyces.
REFERENCE   1 (bases 1 to 5028)
AUTHORS     Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
TITLE       Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
JOURNAL     Yeast 10 (11), 1503-1509 (1994)
MEDLINE     95176709
REFERENCE   2 (bases 1 to 5028)
AUTHORS     Roemer,T., Madden,K., Chang,J. and Snyder,M.
TITLE       Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
JOURNAL     Genes Dev. 10 (7), 777-793 (1996)
MEDLINE     96194260
REFERENCE   3 (bases 1 to 5028)

```

Locus Name

Unique

✦ Up to 10 characters

✦ 6 character

◆ Genus species

✦ 8 characters

◆ Just accession number

✦ Better to search for accession number than Locus Name

The screenshot shows a web browser window with the address bar displaying <http://www.ncbi.nlm.nih.gov/Sitemap/sampleRecord.html#GenBankDivisionB>. The page title is "NCBI Sample GenBank Record". Below the title, there are navigation links: PubMed, Entrez, BLAST, OMIM, Taxonomy, and Structure. The main content area is titled "GenBank Flat File Format" and includes a paragraph explaining that clicking on any link in the sample record will lead to a detailed description of that data element or field. Below this, the GenBank flat file format is displayed, showing fields such as LOCUS, DEFINITION, ACCESSION, VERSION, KEYWORDS, SOURCE, ORGANISM, REFERENCE, and JOURNAL. The LOCUS field shows SCU49845, 5028 bp, DNA, PLN, and 21-JUN-1999. The DEFINITION field shows Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p (AXL2) and Rev7p (REV7) genes, complete cds. The ACCESSION field shows U49845. The VERSION field shows U49845.1 and GI:1293613. The KEYWORDS field is empty. The SOURCE field shows baker's yeast. The ORGANISM field shows Saccharomyces cerevisiae, Eukaryota; Fungi; Ascomycota; Hemiascomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces. The REFERENCE field shows three references: 1 (bases 1 to 5028) by Torpey, L.E., Gibbs, P.E., Nelson, J. and Lawrence, C.W. (1994), 2 (bases 1 to 5028) by Roemer, T., Madden, K., Chang, J. and Snyder, M. (1996), and 3 (bases 1 to 5028) by Genes Dev. 10 (7), 777-793 (1996). The JOURNAL field shows Yeast 10 (11), 1503-1509 (1994) for reference 1, and Genes Dev. 10 (7), 777-793 (1996) for reference 2.

NCBI Sample GenBank Record

PubMed Entrez BLAST OMIM Taxonomy Structure

GenBank Flat File Format

Click on any link in this sample record to see a detailed description of that data element or field. All of the descriptions are included on this page, so it can be printed as a single document. You can also return to the [Alphabetical Quicklinks Table](#) or [Resource Guide](#)

LOCUS SCU49845 5028 bp DNA PLN 21-JUN-1999

DEFINITION Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p (AXL2) and Rev7p (REV7) genes, complete cds.

ACCESSION U49845

VERSION U49845.1 GI:1293613

KEYWORDS

SOURCE baker's yeast.

ORGANISM Saccharomyces cerevisiae
Eukaryota; Fungi; Ascomycota; Hemiascomycetes; Saccharomycetales;
Saccharomycetaceae; Saccharomyces.

REFERENCE 1 (bases 1 to 5028)

AUTHORS Torpey, L.E., Gibbs, P.E., Nelson, J. and Lawrence, C.W.

TITLE Cloning and sequence of REV7, a gene whose function is required for DNA damage-induced mutagenesis in Saccharomyces cerevisiae

JOURNAL Yeast 10 (11), 1503-1509 (1994)

MEDLINE 95176709

REFERENCE 2 (bases 1 to 5028)

AUTHORS Roemer, T., Madden, K., Chang, J. and Snyder, M.

TITLE Selection of axial growth sites in yeast requires Axl2p, a novel plasma membrane glycoprotein

JOURNAL Genes Dev. 10 (7), 777-793 (1996)

MEDLINE 96194260

REFERENCE 3 (bases 1 to 5028)

AUTHORS

TITLE

JOURNAL

MEDLINE

REFERENCE

GenBank Accession Number

- ✧ Unique identifier for sequence record
- ✧ Usually a combination of letter(s) and numbers
- ✧ Do not change even if information changes
- ✧ Newer accession numbers to new submission using some of this data

The screenshot shows a web browser window displaying a GenBank record. The address bar shows the URL: <http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html>. The page title is "Haven, CT, USA". The "FEATURES" section is highlighted with a red circle. The "CDS" (Coding Sequence) feature is also highlighted with a red circle. The "gene" feature is highlighted with a red circle. The "translation" field shows the amino acid sequence: "MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF...".

FEATURES

Location/Qualifiers

1..5028

/organism="Saccharomyces cerevisiae"

/db_xref="taxon:4932"

/chromosome="IX"

/map="9"

/codon_start=3

/product="TCP1-beta"

/protein_id="AAA98665.1"

/db_xref="GI:1293614"

/translation="SSIYNGISTSGLDLNNGTIADMRLGIVESYKLRVSSASEA
AEVLLRVNIIIRARPRTANRQH"

687..3158

/gene="AXL2"

687..3158

/gene="AXL2"

/note="plasma membrane glycoprotein"

/codon_start=1

/function="required for axial budding pattern of S.
cerevisiae"

/product="Ax12p"

/protein_id="AAA98666.1"

/db_xref="GI:1293615"

/translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF
TFQISNDTYKSSVDKTAQITTYNCFDLPWLSFDSSSRTPSGEPSSDLLSDANTLYFN
VILEGTDADSTSLNNTYQFVWYNRPSISLSSDFNLLALLKNYGVYNGKNAKLDPNE
VFNVTDPDSMTNEESIYSGYRSQLYNAPLPNULFFDSGELKFTGTAPVINSALAPE
TSYSFVLIATDIEGFSAVEVEFELVIGAHQLTTSIQNSLLINVTDTGNVSYDLPLNV
YLDLDPISDPLGSLINLLDAPDWALDNATISGSVPDELLGKNSNPANFSVSIYDITYG
DVIYFNFEVYSTDLFAISSLPNINATRGWFSYTFLPSTQFTDVTNINVSLEFNTSSQ
DHDVVKFQSSNLTLAGVPHKFDKLSGLKANQGSQSELYFNIIGHDSKITHNSHA
NATSTRSSHHSSTSTSYSTSTYTTAKISSTSAATSSAPALPANKTSSHNKKAVALA
CGVAIPLGVILVALICFLIFRRRRRNPDPDENLPHASOPDLNPNANKPNQENATPLN
NPFDDASSYDDTIAARLAALNTLKLNDHSAATESISSVDEKRDLSGCHNTYNDQFQ
SQSKELLAKPPVQPPSPFFDPQNRSSSVYMDSEPAVKNKSWRYTGKLSPVSDIVRDS
YGSQKTVDTKFLFDLEAPEKEKRTSRDVTMSSLDPWNSNISPSPVKSVTPSPYNTK

Address <http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html> Go Links

gene
CDS

```

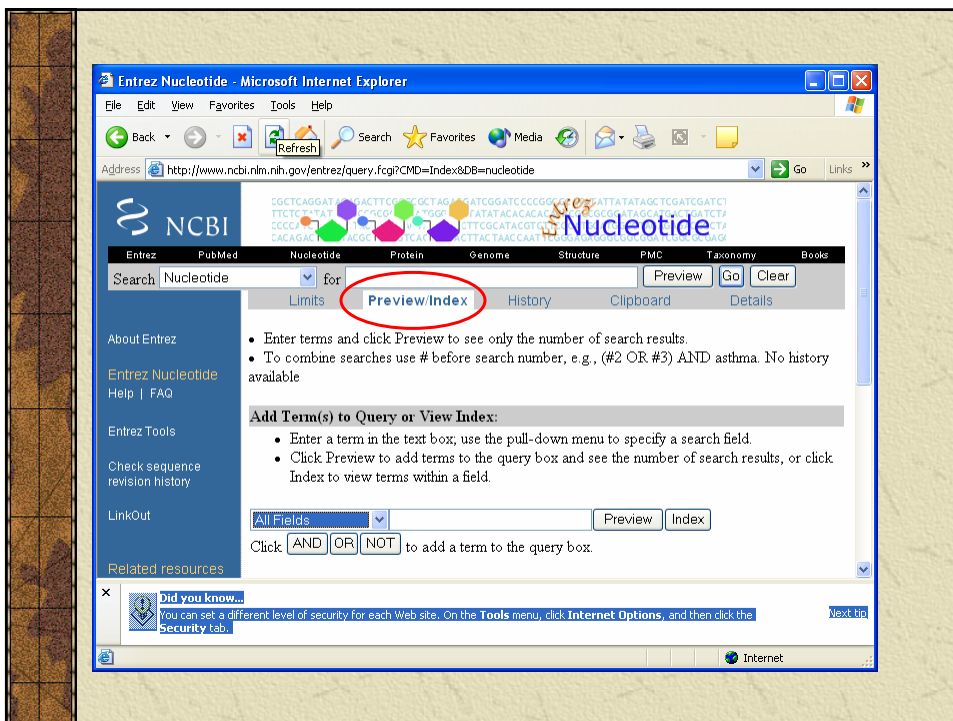
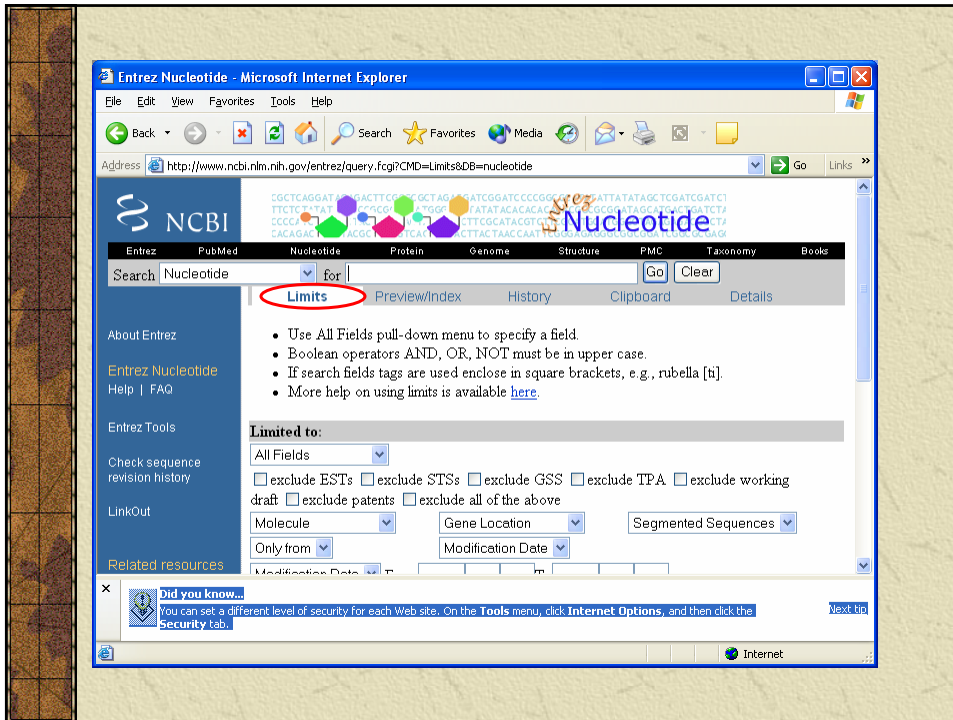
NATSTRSSHHSTSTSSYTSSTYAKISSTSSAAATSSAPAAPAANKTSSHNNKAVAIA
CGVAIPLGVILVALICFLIFWRRRRENPDDENLPHAISGPDNNPANKPNQENATPLN
NPFDDDDASSYDDTSIARRLAALNTLKLNDHSAATESD ISSVDEKRDSLGMNTYNDQFQ
SQSKEELLAKFPVPQPPESPFFDPQNRSSSVYMDSEPAVNKSURYTGNSPVS DIVRDS
YGSQKTVDTKFLDLEAPEKEKRTSRDVTMSSSLDPWNSNISPSVPKSVTPSPYNTK
HNRRLQNIQDSQSGKNGITPTMTSTSSDDFVPVKDGENFCVWHSMEPDRRPSKKRL
VDFSNNKSNVNVGVQKDIHGRIPEML"
complement (3300..4037)
/gene="REV7"
CDS
complement (3300..4037)
/gene="REV7"
/codon_start=1
/product="Rev7p"
/protein_id="AAA98667.1"
/db_xref="GI:1293616"
/translation="MNRWVEKWLRVYLKCYINLILFYRNVYPQSFYTTYQSFNLPQ
FVPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKNDLCIEKYVLDVDFSELQHV
KDDQIITETEVDFEFRSSLNLSLHLEKLPKVNDDTITFEAVINAIELELGHKLDNR
RVDLSLEEKAEIERDSNWKQEDENLPDNNGFQPPRIKLTSLVGSVDVGLIHHQFSEK
LISGDDKILNGVYSQYEEGESIFGSLF"
BASE COUNT      1510 a   1074 c   835 g   1609 t
ORIGIN
1 gatcctccat atacaacggt atctccacgt caggtttaga tctcaacaac ggaaccattg
61 ccgacatgag acagttaggt atcgctcgaga gttacaagct aaaacgagca gtagtcagct
121 ctgcactctga agccgctgaa gttctactaa ggggtggataa catcatccgt gcaagaccaa
181 gaaccgcgcaa tagacaacat atgtaacata tttaggatat accctgaaaa taataaacgg
241 ccacactgtc attattataa ttagaacacg aacgcaaaaa ttatcccata tataattcaa
301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa
361 attttggcaa cttatgttct ctcttcgagc agtaactcgag ccctgtctca agaattgaat
421 aatacccatc gttaggtatgg ttaagatag catctccaca acctcaaaag tccttgccga
481 gagtcgcctc cctttgtcga gtaattttca cttttcatat gagaacttat ttctttatc
541 tttactctca catcctgtag tgattgacac tgcacacagcc accatcacta gaagaacaga
601 acaattactt aatagaaaaa ttatatcttc ctgaaaacga ttctctgctt ccaacactca
661 cgtatatcaa gaagcattca cttaccatga cacagcttca gatttcatta ttgctgacga
721 ctactatata cactactcat ctagtatggg ccacgccccta tgaggcatac cctatcgga
781 aacaataccc ccagtgagca agagtcgaat aatcgtttac atttcaaat tccaatgata
841 cctataaatc gtctgtagac aagacagctc aaataacata caattgcttc gacttaccga
901 gctggcttct gtttgactct agttctagaa cgttctcagg tgaacctctt tctgacttac

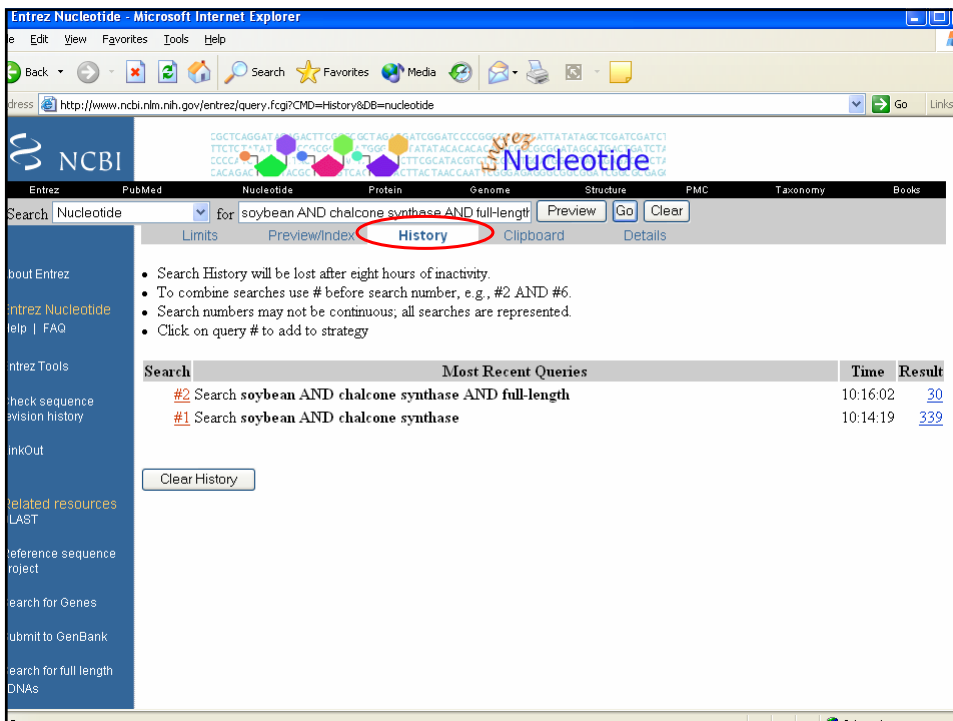
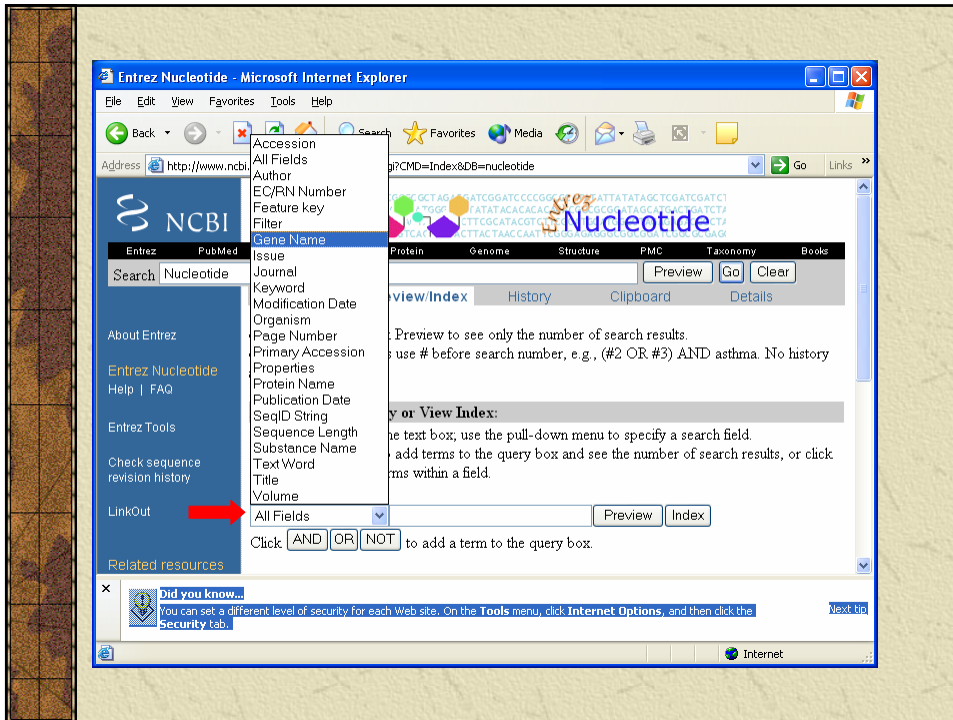
```

Internet

Refine the Query

- ✧ Often a search finds too many (or too few) sequences, so you can go back and try again with more (or fewer) keywords in your query
- ✧ The “**History**” feature allows you to combine any of your past queries.
- ✧ The “**Limits**” feature allows you to limit a query to specific organisms, sequences submitted during a specific period of time, etc.
- ✧ [Many other features are designed to search for literature in MEDLINE]





DNA similarity search



Find related sequences

- ✚ Find ESTs
 - ◆ If not full-length, may allow assembly from ESTs
- ✚ Find other family members
 - ◆ Organization and function
- ✚ Find similar genes from other cultivars
 - ◆ SNP discovery
- ✚ Find similar genes from other organisms
 - ◆ Phylogenetic relationships

ESTs (Expressed Sequence Tags)

- ✧ partial cDNA sequences
- ✧ dbEST at NCBI
 - ◆ a comprehensive set of all public EST data
- ✧ UniGene at NCBI
 - ◆ clusters of ESTs and known genes from key species
 - does NOT have consensus sequences
 - has far too many clusters to be representative of real genes (129 K human clusters)

Find related DNA sequences

- ✧ Similarity Search (BLAST)
- ✧ NCBI GenBank database

BLAST Searches

- ✧ Compare your sequence with database
- ✧ <http://www.ncbi.nlm.nih.gov/BLAST/>
- ✧ Nucleotide
- ✧ Protein
- ✧ Targeted to a genome

✧ BLAST

- ◆ **Basic Local Alignment Search Tool**
- ◆ Local alignment
- ◆ Tutorial at:
<http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html>

BLAST

Discontiguous Mega BLAST

- ◆ Comparison of diverged sequences especially from different organisms
- ◆ Alignments with low degree of identity
- ◆ Looks for hits in “non-consecutive positions”

Mega BLAST

- ◆ Slight differences in similarity
- ◆ Not effective at low degree of identity
- ◆ Faster; handles longer sequences

BLAST

- ◆ Local alignment tool
- ◆ <http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html>

The screenshot shows the NCBI BLAST website interface. At the top, there's a navigation bar with links to PubMed, Entrez, BLAST, OMIM, Taxonomy, and Structure. Below this, a banner for BLAST 2.2.9 is displayed, dated 12 May 2004. The main content area is divided into several sections: Nucleotide, Protein, Translated, Genomes, Special, and Meta. Each section contains a list of links to various BLAST tools and databases. On the left side, there's a sidebar with links to FAQs, News, References, NCBI Contributors, Education (Program selection guide, Tutorial, URL API guide), Download (Databases, Documentation, Executables, Source code), and Support (Helpdesk, Mailing list). At the bottom, there's a security warning from Internet Explorer.

File Edit View Favorites Tools Help

Back Forward Stop Reload Home Search Favorites Media Print Mail Link

Address <http://www.ncbi.nlm.nih.gov/BLAST/> Go Links

NCBI BLAST

PubMed Entrez BLAST OMIM Taxonomy Structure

Info

NEW 12 May 2004 BLAST 2.2.9 has been released. [Read more...](#)

Nucleotide

- Discontiguous megablast
- Megablast
- Nucleotide-nucleotide BLAST (blastn)
- Search for short, nearly exact matches
- Search trace archives with megablast or discontiguous megablast

Protein

- Protein-protein BLAST (blastp)
- PHI- and PSI-BLAST
- Search for short, nearly exact matches
- Search the conserved domain database (rpsblast)
- Search by domain architecture (cdart)

Translated

- Translated query vs. protein database (blastx)
- Protein query vs. translated database (blastn)
- Translated query vs. translated database (blastx)

Genomes

- Chicken, cow, pig, dog, sheep, cat **NEW**
- Environmental samples
- Human, mouse, rat
- Fugu rubripes, zebrafish
- Insects, nematodes, plants, fungi, malaria
- Microbial genomes, other eukaryotic genomes

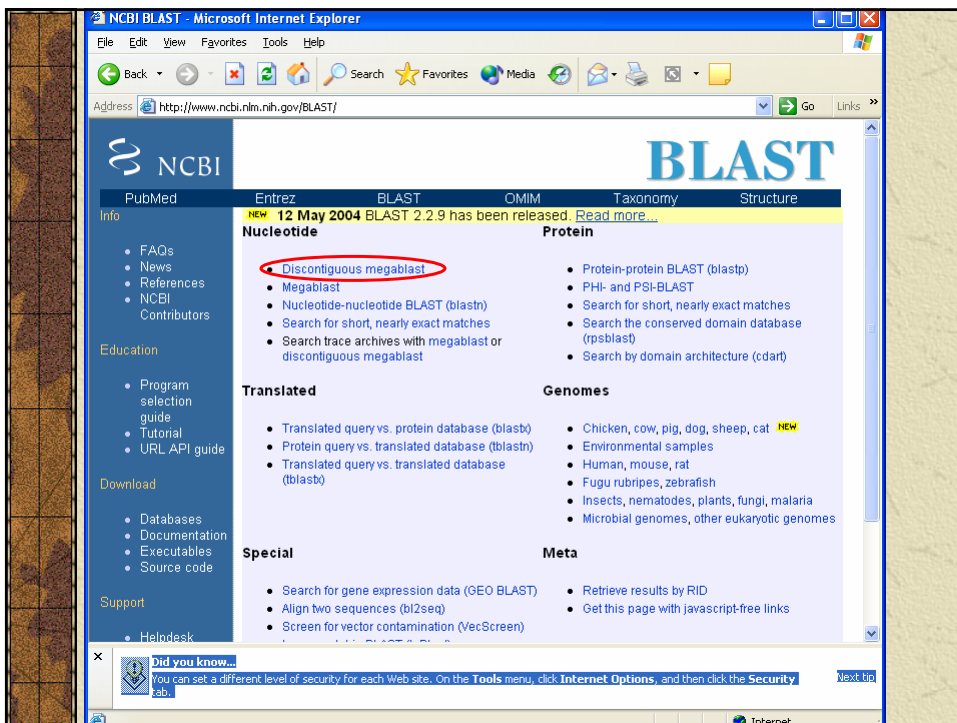
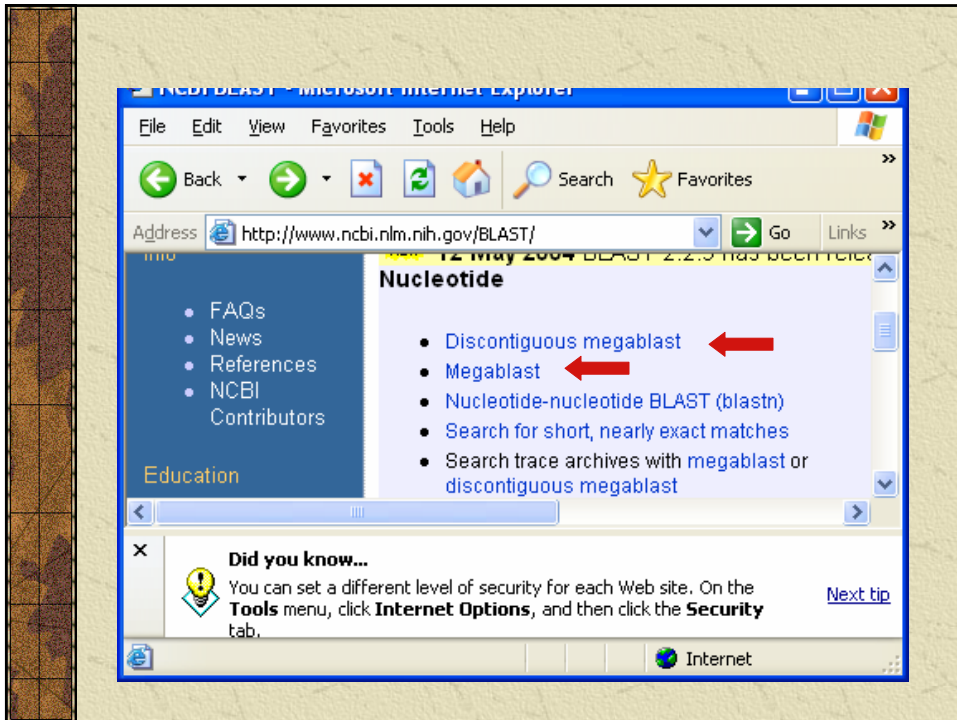
Special

- Search for gene expression data (GEO BLAST)
- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobulin BLAST (IgBlast)

Meta

- Retrieve results by RID
- Get this page with javascript-free links

Did you know...
You can set a different level of security for each Web site. On the **Tools** menu, click **Internet Options**, and then click the **Security** tab. [Next tip](#)



You have a sequence.
Does it have similarity to other known genes???
Copy DNA sequence from file

```
GGTTGTGCCAAGTCCATTTTCTATTGACTTCTTCCTCATTTGATCCAAGATGAACAGCAC
ACATGCACTTGACATGTTACCATACTCGCTAAGCACGTGTCTAGTAGCTTCCATTTTTTC
ATGCTTCAATCCTAACTTAGCCTCAACTTGGTCCAAAATTGCTGGTCCACCAGGGTGTGC
AATCCAAAAGATAGAGTTGTAATCATCAATTTCTAAGGGTTTGAAGGCTTCAACCAAGGC
CTTTTCGATGTTCTTGGAGATGAGTCCAGGAACATCCTTGAGGAGATGGAAAGTGAGTCC
TACTTGGCGAAGGTGGCCATCAATAGCGCCTTCGCTGTCTGGAAGGATTGTTTGTGCAGT
CCACACAAGCTCAAAACAAGGCTTTTCAGCTGGCAGAGGATCTGATCCAACAATGACAGC
GCTGCACCATCTCCAAACAAGGCTTGCCCCACAAGGCTGTCAAGATGTGTGTCACTCGG

GCCACGAAATGTGACTGCTGTGATCTCCGA
```

NCBI Blast - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Search Favorites Media Print Mail Print Mail Print

Address http://www.ncbi.nlm.nih.gov/BLAST/blast.cgi?ALIGNMENTS=50&ALIGNMENT_VIEW=Pairwise&AUTO_FORMAT=Semiauto& Go Links

NCBI megablast BLAST

Nucleotide Protein Translations Retrieve results for an RLD

What is discontinuous Mega BLAST?

Search

Load query file from disk: Browse...

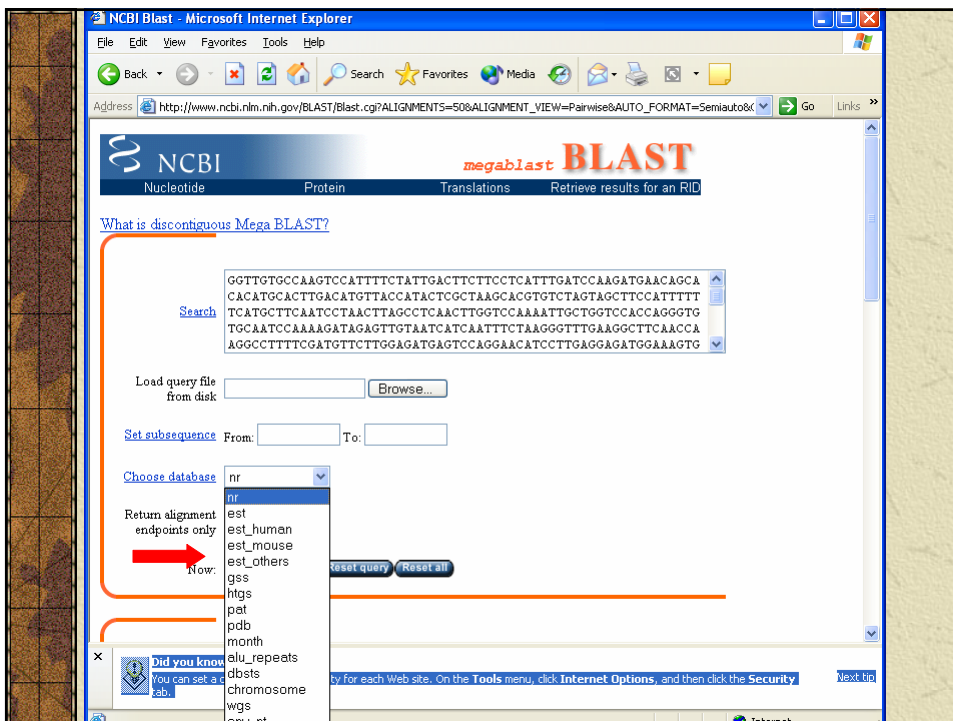
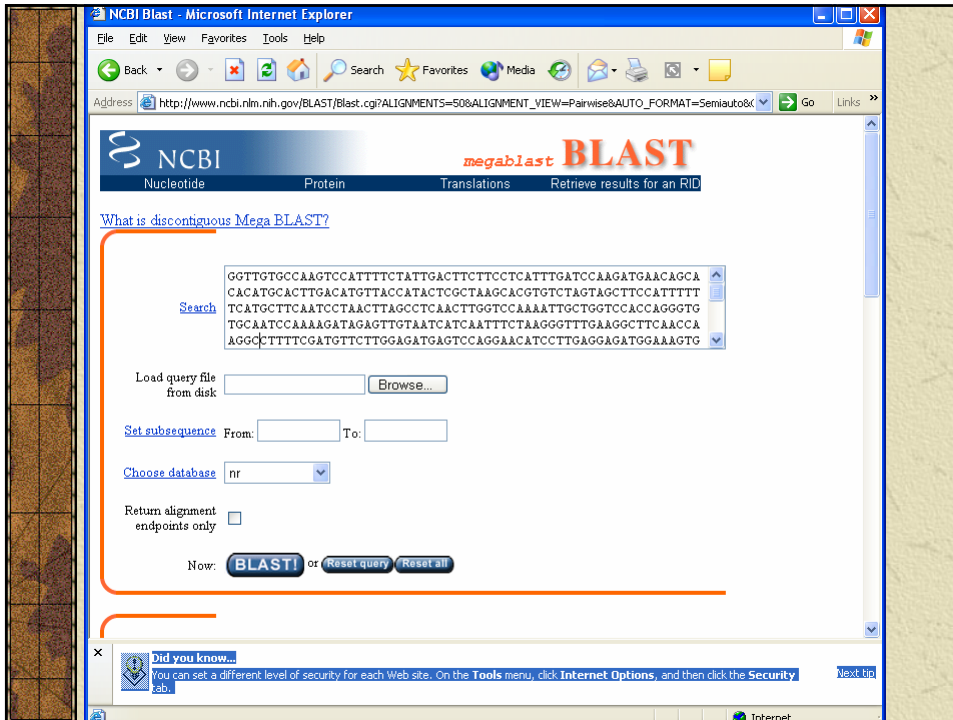
Set subsequence From: To:

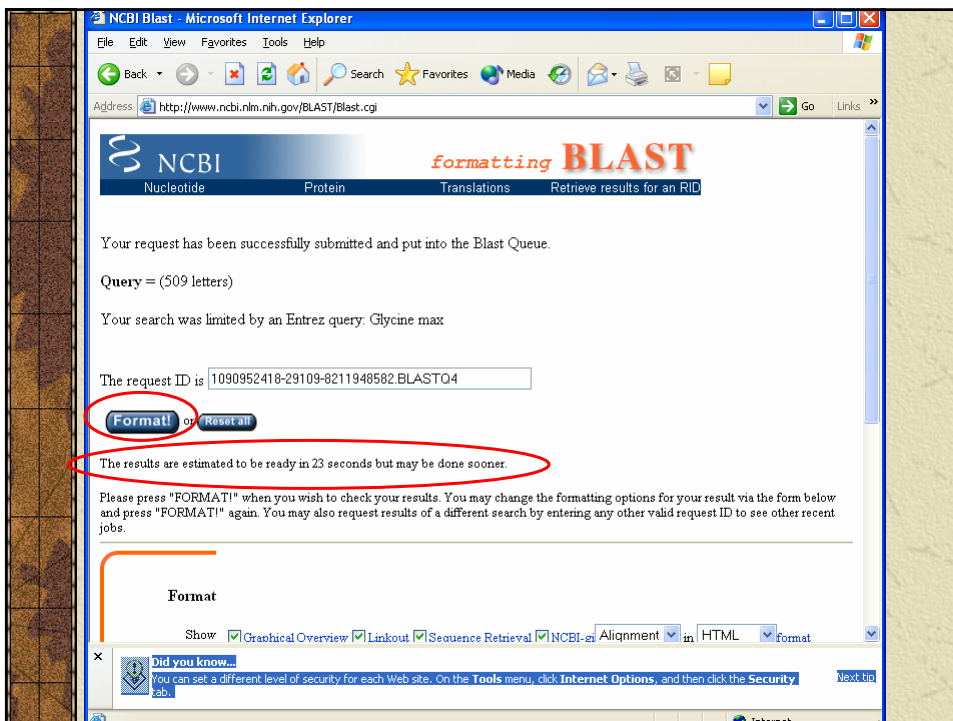
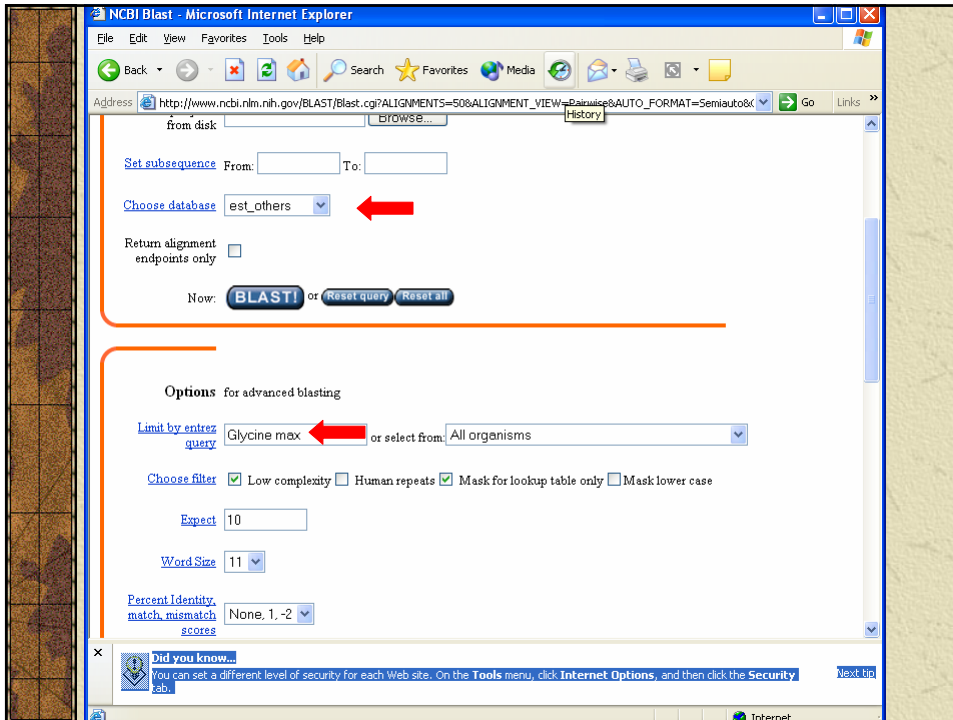
Choose database:


Return alignment endpoints only ☐

Now: **BLAST!** or **Reset query** **Reset all**

Did you know... You can set a different level of security for each Web site. On the Tools menu, click Internet Options, and then click the Security tab. Next bio







formatting **BLAST**

[Nucleotide](#)
[Protein](#)
[Translations](#)
[Retrieve results for an RID](#)

Your request has been successfully submitted and put into the Blast Queue.
Query = (1509 letters)
 Your search was limited by an Entrez query: Glycine max



The request ID is


Format! or **Reset all**

The results are estimated to be ready in 37 seconds but may be done sooner.
 Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media

Address Go Links


results of **BLAST**

BLASTN 2.2.9 [May-01-2004]

RID: 1090952418-29109-8211948582.BLASTQ4

Database: GenBank non-mouse and non-human EST entries
 12,989,815 sequences; 6,991,256,704 total letters

If you have any problems or questions with the results of this search
 please refer to the [BLAST FAQ](#)

[Taxonomy reports](#)


Query=
 (509 letters)


[Distribution of 101 Blast Hits on the Query Sequence](#)

Mouse-over to show define and scores. Click to show alignments

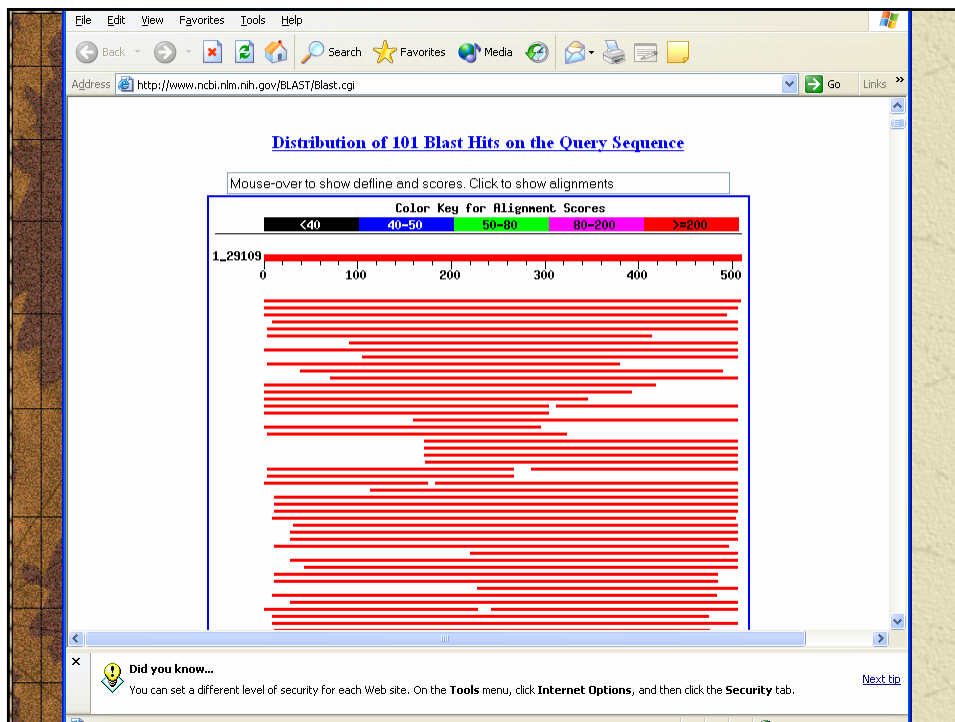
Color Key for Alignment Scores

<40	40-50	50-80	80-200	>=200
-----	-------	-------	--------	-------

1_29109 


Did you know...
 You can set a different level of security for each Web site. On the **Tools** menu, click **Internet Options**, and then click the **Security** tab.

[Next tip](#)



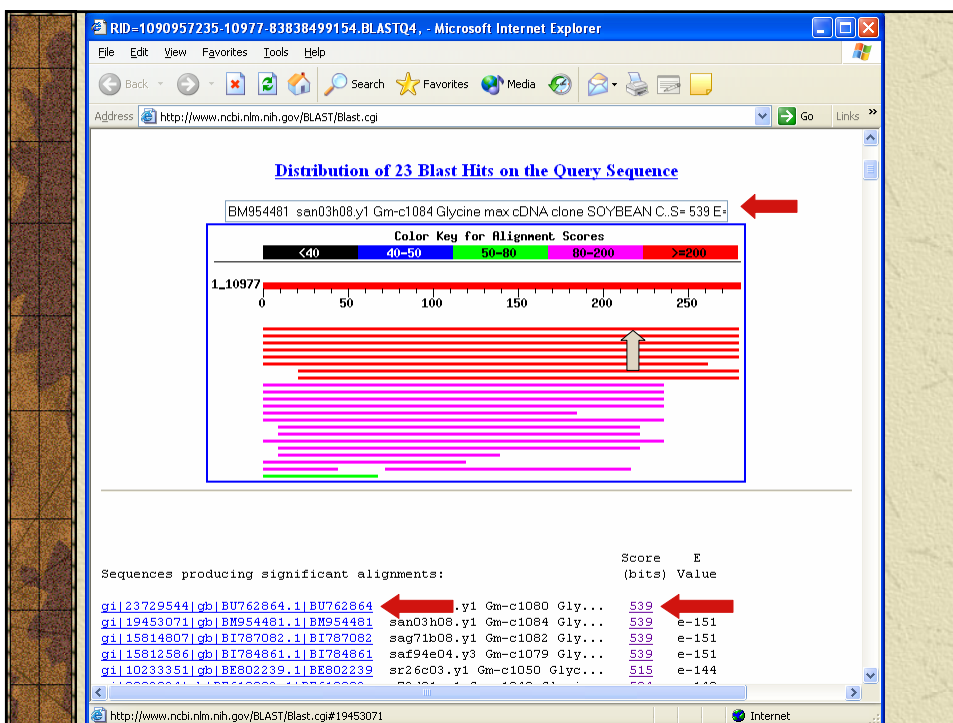
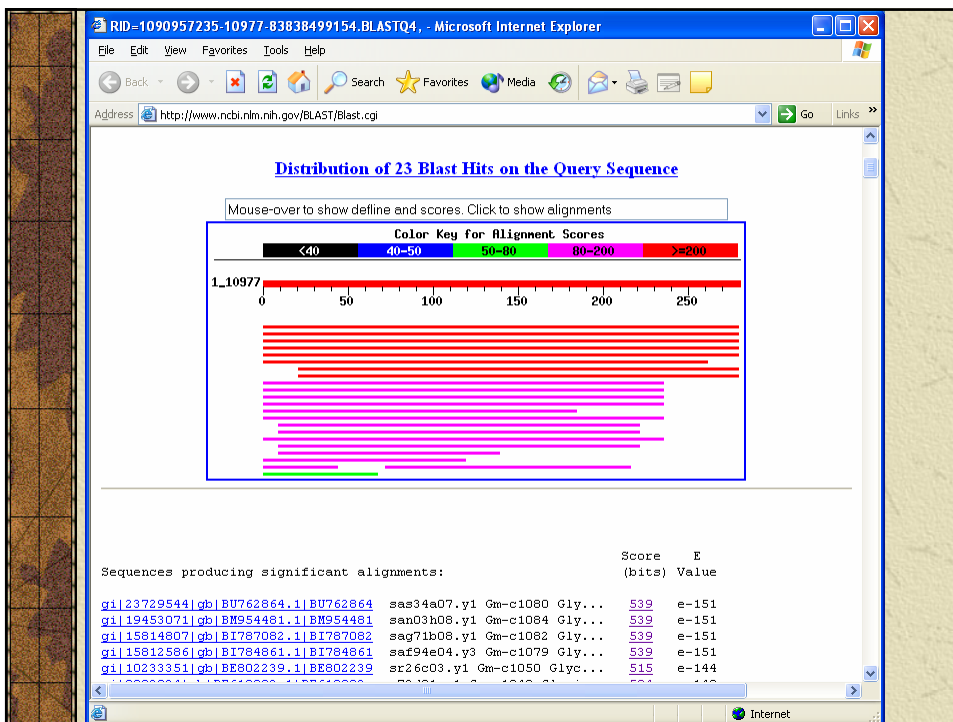
File Edit View Favorites Tools Help

Address <http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi>

Sequences producing significant alignments: (bits) Value

gi 14008724 gb BG725328.1 BG725328	sae35d12.y1 Gm-c1051 Gly...	965	0.0	U
gi 27427571 gb CA939091.1 CA939091	sav41g09.y1 Gm-c1069 Gly...	712	0.0	U
gi 16346573 gb BI972168.1 BI972168	sag88a12.y1 Gm-c1084 Gly...	694	0.0	U
gi 15287478 gb BI471369.1 BI471369	sag19f06.y1 Gm-c1080 Gly...	694	0.0	U
gi 37995974 gb CF807563.1 CF807563	psHB025x009f USDA-IFAFS:...	687	0.0	U
gi 23732296 gb BU764307.1 BU764307	sar98d05.y2 Gm-c1080 Gly...	675	0.0	U
gi 37994162 gb CF805908.1 CF805908	psHB001x006f USDA-IFAFS:...	664	0.0	U
gi 16342613 gb BI968208.1 BI968208	GM830004B12F05 Gm-r1083 ...	658	0.0	U
gi 19934725 gb BQ079755.1 BQ079755	san17h09.y1 Gm-c1084 Gly...	642	0.0	U
gi 19934445 gb BQ079475.1 BQ079475	san14c01.y1 Gm-c1084 Gly...	625	e-176	U
gi 13478388 gb BG507884.1 BG507884	sac82e09.y1 Gm-c1072 Gly...	623	e-176	U
gi 37996764 gb CF808353.1 CF808353	psHB034xJ14f USDA-IFAFS:...	592	e-166	U
gi 16106094 gb BI893834.1 BI893834	sag93f11.y1 Gm-c1084 Gly...	587	e-165	U
gi 19935253 gb BQ080283.1 BQ080283	san31a09.y1 Gm-c1084 Gly...	565	e-158	U
gi 11412994 gb BF425005.1 BF425005	su53b09.y1 Gm-c1069 Glyc...	496	e-138	U
gi 37996666 gb CF808255.1 CF808255	psHB033xIO1f USDA-IFAFS:...	465	e-128	U
gi 19934418 gb BQ079448.1 BQ079448	san13h02.y1 Gm-c1084 Gly...	465	e-128	U
gi 37996212 gb CF807801.1 CF807801	psHB028xG08f USDA-IFAFS:...	450	e-124	U
gi 13562975 gb BG551195.1 BG551195	sad34d04.y1 Gm-c1074 Gly...	448	e-123	U
gi 7673732 gb AW160139.1 AW160139	pb1t17 soybean, century c...	442	e-121	U
gi 15815451 gb BI787726.1 BI787726	sag75a07.y1 Gm-c1084 Gly...	435	e-119	U
gi 13480079 gb BG509422.1 BG509422	sad13f03.y1 Gm-c1074 Gly...	435	e-119	U
gi 37996627 gb CF808216.1 CF808216	psHB033xC14f USDA-IFAFS:...	429	e-117	U
gi 23728449 gb BU762277.1 BU762277	sar87d02.y1 Gm-c1074 Gly...	421	e-115	U
gi 15811812 gb BE917591.1 BE917591	GmCHS6 soybean root subt...	419	e-114	U

Did you know... You can set a different level of security for each Web site. On the **Tools** menu, click **Internet Options**, and then click the **Security** tab. [Next tip](#)



FASTA/BLAST Statistics

- ✱ E() value is equivalent to standard P value
- ✱ Significant if $E() < 0.05$ (smaller numbers are more significant)
 - ◆ The E-value represents the likelihood that the observed alignment is due to chance alone. A value of 1 indicates that an alignment this good would happen by chance with any random sequence searched against this database.
- ✱ The histogram should follow expectations (asterisks) except for hits

Interpretation of output

- ✱ very low E() values (e-100) are homologs or identical genes
- ✱ moderate E() values are related genes
- ✱ long list of gradually declining of E() values indicates a large gene family
- ✱ long regions of moderate similarity are more significant than short regions of high identity

What this does for you

- ✧ You identified what gene is encoded by your clone's sequence
- ✧ Perhaps you may have found the function of your gene
- ✧ You have more cDNA sequences to add together to build a consensus and perhaps a full-length cDNA

Biological Relevance

- ✧ It is up to you, the biologist to scrutinize these alignments and determine if they are significant.
- ✧ Were you looking for a short region of nearly identical sequence or a larger region of general similarity?
- ✧ Are the mismatches conservative ones?
- ✧ Are the matching regions important structural components of the genes or just introns and flanking regions?

Borderline similarity

- ✠ What to do with matches with E() values in the 0.5 -1.0 range?
- ✠ this is the “**Twilight Zone**”
- ✠ retest these sequences and look for related hits (not just your original query sequence)
- ✠ similarity is transitive:
if **A~B** and **B~C**, then **A~C**

Advanced Similarity Techniques

Automated ways of using the results of one search to initiate multiple searches

- ✠ **INCA** (Iterative Neighborhood Cluster Aalysis)
<http://itsa.ucsf.edu/~gram/home/inca/>
 - ◆ Takes results of one **BLAST** search, does new searches with each one, then combines all results into a single list
 - ◆ JAVA applet, compatibility problems on some computers
- ✠ **PSI BLAST**
<http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/psi1.html>
 - ◆ Creates a “position specific scoring matrix” from the results of one **BLAST** search
 - ◆ Uses this matrix to do another search
 - ◆ builds a family of related sequences
 - ◆ can’t trust the resulting e-values

File Edit View Favorites Tools Help

Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Nucleotide&list_uids=14008724&dopt=GenBank Go Links

NCBI Nucleotide

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search Nucleotide for Go Clear

Limits Preview/Index History Clipboard Details

Display GenBank Show: 20 Send to File Features Links

☐ 1: BG725328 sae35d12.y1 Gm-cl...[gi.14008724]

LOCUS BG725328 510 bp mRNA linear EST 22-JUL-2004
 DEFINITION sae35d12.y1 Gm-c1051 Glycine max cDNA clone GENOME SYSTEMS CLONE
 ID: Gm-c1051-7103 5' similar to SW:CHS6_SOYEN P30080 CHALCONE
 SYNTHASE 6 ;, mRNA sequence.

ACCESSION BG725328
 VERSION BG725328.1 GI:14008724
 KEYWORDS EST.
 SOURCE Glycine max (soybean)
 ORGANISM [Glycine max](#)
 Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
 Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots; rosids
 ; eurosids I; Fabales; Fabaceae; Papilionoideae; Phaseoleae;
 Glycine.

REFERENCE 1 (bases 1 to 510)
 AUTHORS Shoemaker,R., Keim,P., Vodkin,L., Erpelding,J., Coryell,V., Khanna
 ,A., Bolla,B., Harra,M., Hillier,L., Kucaba,T., Martin,J., Beck,C.,
 Wylie,T., Underwood,K., Steptoe,M., Theising,B., Allen,M., Bowers
 ,Y., Person,B., Swaller,T., Gibbons,M., Pape,D., Harvey,N., Schurk
 ,R., Ritter,E., Kohn,S., Shin,T., Jackson,Y., Cardenas,M., McCann
 ,R., Waterston,R. and Wilson,R.
 TITLE Public Soybean EST Project
 JOURNAL Unpublished (1999)

Did you know...
 You can set a different level of security for each Web site. On the **Tools** menu, click **Internet Options**, and then click the **Security** tab. [Next tip](#)

File Edit View Favorites Tools Help

Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Nucleotide&list_uids=14008724&dopt=GenBank Go Links

/db_xref="taxon:3847"
 /clone="GENOME SYSTEMS CLONE ID: Gm-c1051-7103"
 /tissue_type="floral meristematic mRNA"
 /lab_host="DH10B"
 /clone_lib="Gm-c1051"
 /note="Vector: pBluescript II SK+; Site 1: EcoRI; Site 2:
 XhoI; The cDNA library was constructed from floral
 meristematic mRNA provided by Dr. Halina Knap of Clemson
 University. Complementary DNA was synthesized from mRNA
 using a primer consisting of a poly(dT) sequence with a
 XhoI restriction site. EcoRI adapters were ligated to the
 blunt-ended cDNA fragments followed by XhoI digestion. The
 cDNA fragments were directionally cloned into the
 EcoRI-XhoI restriction site of the pBluescript vector. The
 ligated cDNA fragments were transformed into DH10B host
 cells (GibcoBRL). This library was constructed in the
 laboratory of Dr. Randy Shoemaker."

ORIGIN
 1 gggtgtgcca agtcacatttt ctattgactt cttctcatt tgatccaaga tgaacagcac
 61 acatgcactt gacatgttac catactogct aagcacgtgt ctagtgcgtt ccattttttc
 121 atgttcaat cctaacttgc ctccaacttg gtccaaaatt gctgttccac cagggtgtgc
 181 aatcacaagg atagagttgt aatcatcaat ttctaaaggt ttgaaggctt caaccaaggc
 241 ctttttgatg ttcttgagga tgaagtccag acatcccttg agggatagga aagtgaagtc
 301 tacttgagga aggtggocat caatagcgcc ttgcgtgtct ggaaggattt ttgtgcagt
 361 caacacaagc tcaacaaagg gcttttcagc tggcagagga ttgatccaa caatgacagc
 421 agctgcacca tctccaaaca aggtttgccc cacaaggctg tcaagatgtg tgctcactgg
 481 gccacgaagt gtgactgctg tgatctccga
 //

[Disclaimer](#) | [Write to the Help Desk](#)
 NCBI | NLM | NIH

Did you know...
 You can set a different level of security for each Web site. On the **Tools** menu, click **Internet Options**, and then click the **Security** tab. [Next tip](#)

FASTA format

- ✚ One of three formats used for sequences
- ✚ Begins with single-line description followed by sequence data

- ✚ Description line starts with ">"

✚ Example:

```
>gij532319|pir|TVFV2E|TVFV2E envelope protein
ELRLRYCAPAGFALLKCNDADYDGFKTNC SNVSVVHCTNLMNTTVTTGLLNGSYSENRT
QIWQKHRTSNDSALILLNKHYNLTVCKRPGNKTVLPVTIMAGLVFHSQKYNLRLRQAWC
HFPSNWKGAWKEVKEEIVNLPKERYRGTDNPKRIFFQRQWGD PETANLWFNCHGEFFYCK
MDWFLNLYNNLTVDADHNECKNTSGTKSGNKRAPGPCVQRTYVACHIRSVIIWLETISKK
TYAPPREGHLECTSTVTGMTVELNYIPKNRTNVTLS PQIESIWAAELDRYKLVEITPIGF
APTEVRRYTGGHERQKRVPFVXXXXXXXXXXXXXXXXXXXXXXXXXVQSQHLLAGILQQQKNL
LAAVEAQQQMLKLTIWGVK
```

The nucleic acid codes supported are:

A → adenosine	M → A C (amino)
C → cytidine	S → G C (strong)
G → guanine	W → A T (weak)
T → thymidine	B → G T C
U → uridine	D → G A T
R → G A (purine)	H → A C T
Y → T C (pyrimidine)	V → G C A
K → G T (keto)	N → A G C T (any)

- gap of indeterminate length

accepted amino acid codes are:

A alanine	P proline
B aspartate or asparagine	Q glutamine
C cystine	R arginine
D aspartate	S serine
E glutamate	T threonine
F phenylalanine	U selenocysteine
G glycine	V valine
H histidine	W tryptophan
I isoleucine	Y tyrosine
K lysine	Z glutamate or glutamine
L leucine	X any
M methionine	* translation stop
N asparagine	

- gap of indeterminate length

Types of data integrated in genome browsers

- Genomic sequence
- RefSeq mRNAs (non-redundant)
- GenBank mRNAs (redundant)
- ESTs
- Gene predictions
- SNPs
- Homologous sequences from other organisms
- STSs

Other Sequence Search Tools

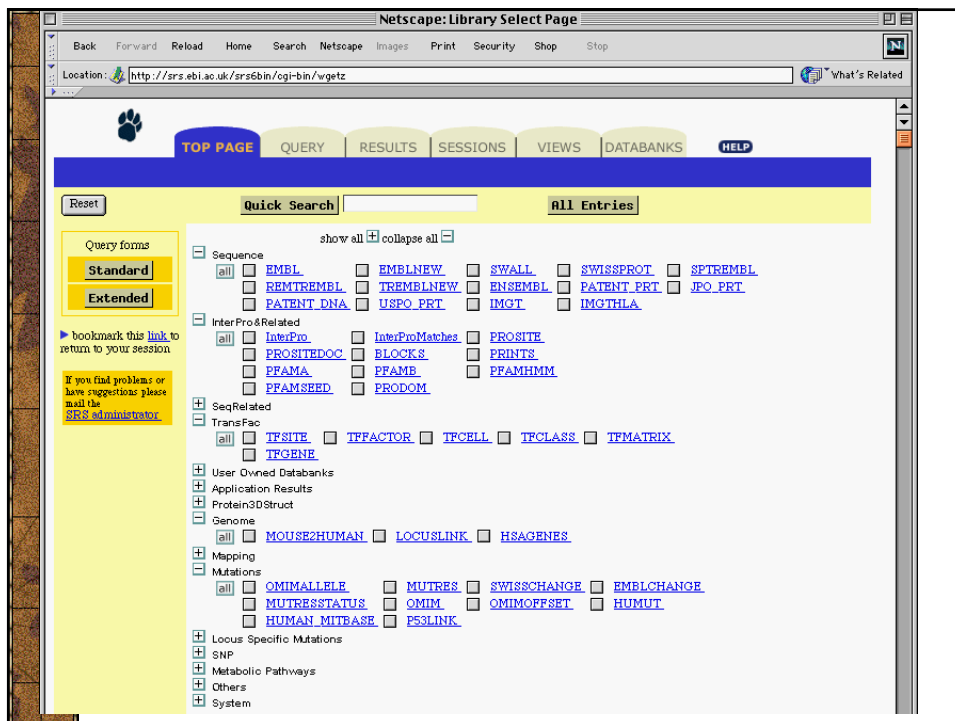
- ✳ **SRS** (Sequence Retrieval Service) was created by Dr. Thure Etzold: *[CABIOS 9(1); 49-57 (1993)]*
- ✳ It is a meta search engine for all types of biological data in hundreds of databases as well as about 20 sequence analysis programs
- ✳ SRS can be accessed over the WWW from many servers (mostly in Europe):

<http://srs.ebi.ac.uk/>

<http://www.infobiogen.fr/srs6bin/cgi-bin/wgetz?-page+top>

<http://www.sanger.ac.uk/srs6bin/cgi-bin/wgetz?-page+top>

<http://iubio.bio.indiana.edu/srs6bin/cgi-bin/wgetz?-page+top>



The screenshot shows a Netscape browser window titled "Netscape: Query Form". The address bar displays the URL `http://srs.ebi.ac.uk/srs6bin/cgi-bin/wgetz`. The browser's menu bar includes Back, Forward, Reload, Home, Search, Netscape, Images, Print, Security, Shop, and Stop. The page features a navigation bar with tabs for TOP PAGE, QUERY (highlighted), RESULTS, SESSIONS, VIEWS, DATABANKS, and HELP. Below the navigation bar, there is a search area with a "Reset" button and a search bar containing the text: `EMBL EMBLNEW SWALL SWISSPROT SPTREMBL REMTREMBL TREMBLNEW ENSEMBL PATENT_PRT JPO_PRT PATENT_DNA USPO_PRT IMGT IMGTHLA`. An "Info" button and a dropdown menu for "about field" (set to "AllText") are also present. A "Submit Query" button is located on the left. Below it, there are options for "append wildcards to words" (checked), "combine searches with" (set to "AND"), and "Number of entries to display per page" (set to "30"). An "Extended query form" button is also visible. The main query area includes a note "separate multiple values by & (and), / (or), /(and not)" and four input fields for "AllText", "AccNumber", "Keywords", and "Organism". Below these is a "retrieve entries of type" dropdown set to "Entry". A "Use predefined view" section shows "* Complete entries *". A "Create your own view" section includes a "Select fields to display" list with options: ID, AccNumber, Description, Keywords, Organism, SeqLength, and Feature: ID. A "sequence format" dropdown is set to "genbank".

Why So Many Databases?

✖ If GenBank has all sequence data and Entrez is such a good query tool, then why are there so many other sequence databases?

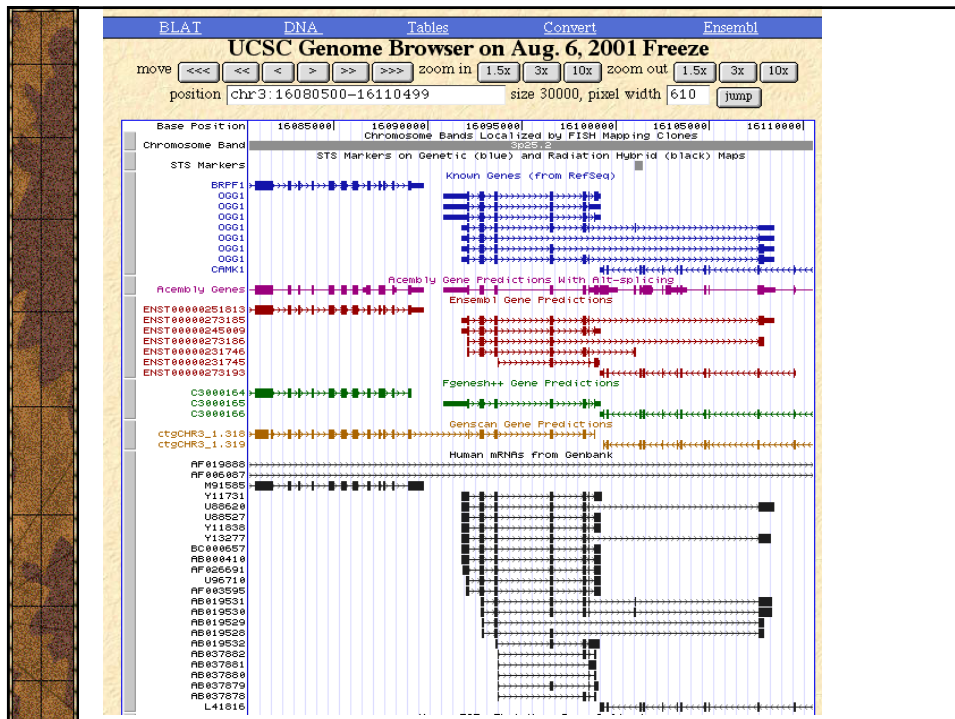
- ◆ Specialized data (single species, immunoglobulins, etc.)
- ◆ Better annotation (i.e. SwissProt)
- ◆ Sequences linked to other data (ACEDB)
- ◆ Subbornness and local pride - EMBL, DDBJ

✖ Well designed databases are interlinked with others for supplemental data

✖ It is very hard to get all relevant information across all databases for any gene

Other Genetic Databases

- ✧ Genome Sequence - where does a gene fall on the genome
 - ◆ integrate multiple layers of information
 - > Sequence contigs, mRNAs, predicted exons, etc.
 - ◆ Single species?
- ✧ ESTs: dbEST @ NCBI
- ✧ SNPs: dbSNP @ NCBI,
 - <http://snp.cshl.org> (SNP Consortium)
- ✧ Metabolism/Pathways
- ✧ Gene Function (Genome Ontology)
- ✧ Protein motifs/domains and protein families



Genome Databases

- ✧ New area - in desperate need of development

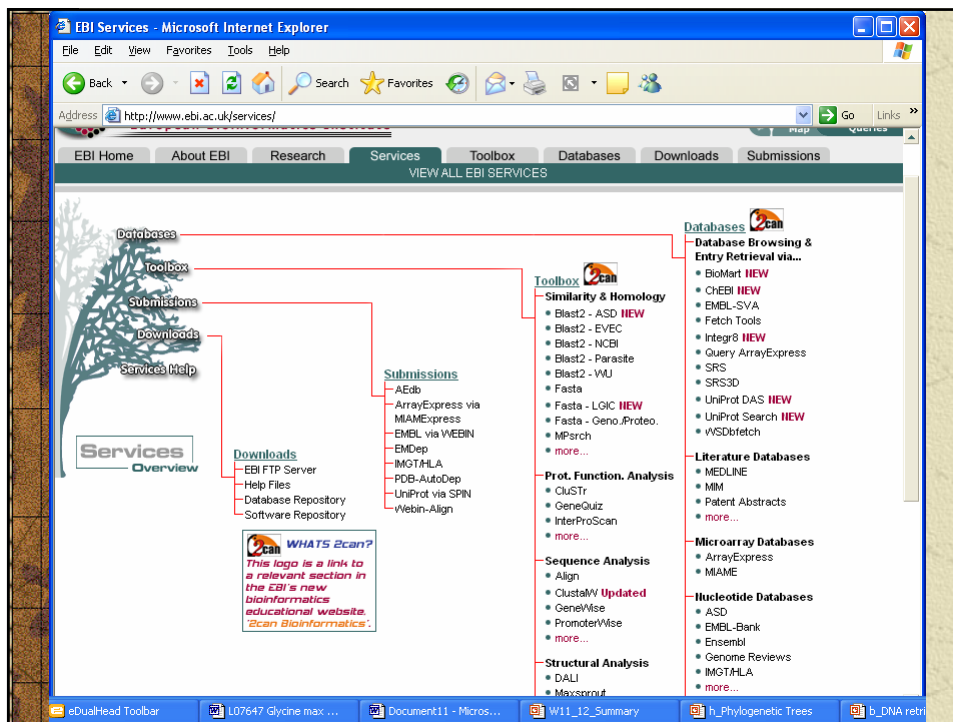
Chromosomes::Sequence::Contigs::Clones::
STS Markers::Genetic Markers::Genes::
Features::Expression data::Phenotype

- ✧ No single database can hold it all
- ✧ UCSC is probably the best right now
genome.ucsc.edu

- ✧ Need a data exchange and linkage infrastructure

European Bioinformatics Institute

- ✧ Products and services
- ✧ Databases
 - ◆ Literature
 - ◆ Microarray
 - ◆ nucleotide
- ✧ Toolbox with software
 - ◆ Similarity searches
 - ◆ Protein function
 - ◆ Sequence analysis
 - ◆ Structure
- ✧ <http://www.ebi.ac.uk/>



IDENTIFIERS	
dbEST Id:	101883
EST name:	yb01a01.s1
GenBank Acc:	T48601
GDB Id:	490761
CLONE INFO	
Clone Id:	IMAGE:69864 (3')
Other ESTs on clone:	yb01a01.r1
DNA type:	cDNA
PRIMERS	
Sequencing:	-21ml3
PolyA Tail:	Unknown
SEQUENCE	
	GGCGGCTCAGTAGCAGGTGCCGTCCACCTCCGCCATGACACAGACACATTGACATGGGT GGGTTTACCACCAAGCGTCCGATGGTCTTCTGTGTGAAGGCCAGCCAGCGCCCTCCATGG CACCATGACAGGAGAAGGNTCCCCCTTCTTCCAGTCTCCGGCTGCCAGCGCGAGTATGCT GGTCACACGAAGGTCGTGGTCCCTGGCTGGNCTCTNCANGGATGCCCAAGTCAGGTACT TNTCGCGGGGACGTCCTGTGACCCCTGCAGCCAGCGAACCAGCAGCTCCTTGGGGCTTN AAGCNGCGCTACCAGGCACTTCAACCGTTCNCCAGCTTCGTTTCAGGGCCACCTTC
Quality:	High quality sequence stops at base: 277
Entry Created:	Feb 6 1995
Last Updated:	Feb 6 1995
COMMENTS	
	High quality sequence stops: 277
	Source: IMAGE Consortium, LLNL
	This clone is available royalty-free through LLNL ; contact the IMAGE Consortium (info@image.llnl.gov) for further information.
PUTATIVE ID	
	Assigned by submitter
	similar to gb:S71043_rnal IG ALPHA-2 CHAIN C REGION (HUMAN)
LIBRARY	
Lib Name:	Stratagene placenta (#937225)
Organism:	<u>Homo sapiens</u>
Sex:	male
Organ:	placenta
Lab host:	SOLR cells (kanamycin resistant)
Vector:	pBluescript SK-
R. Site 1:	EcoRI
R. Site 2:	XhoI
Description:	Cloned unidirectionally. Primer: Oligo dT. Caucasian. Average insert size: 1.2 kb; Uni-ZAP XR Vector; ~5' adaptor sequence: 5' GAAATCGGCAGCAG 3' ~3' adaptor sequence: 5' CTCGAGTATTTTTTTTTTTTTTT 3'

Database Search Strategies

- ✧ General search principles - not limited to sequence (or to biology)
- ✧ Use accession numbers whenever possible
- ✧ Start with broad keywords and narrow the search using more specific terms
- ✧ Try variants of spelling, numbers, etc.
- ✧ Search all relevant databases
- ✧ **Be persistent!!**

What we covered today

- ✧ Retrieving a known DNA sequence
- ✧ Similarity searching with a DNA sequence
- ✧ BLAST